CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	The feature understandability scale for human-centred explainable AI: assessing tabular feature importance
Authors	Rossberg, Nicola;Kleinberg, Bennett;O'Sullivan, Barry;Longo, Luca;Visentin, Andrea
Publication date	2025-10-10
Original Citation	Rossberg, N, Kleinberg, B, O'Sullivan, B, Longo, L & Visentin, A 2025 'The feature understandability scale for human-centred explainable AI: assessing tabular feature importance' CoRR, arXiv, pp. 1-45. https://doi.org/10.48550/arXiv.2510.07050
Type of publication	Article (preprint)
Link to publisher's version	10.48550/arXiv.2510.07050
Rights	© 2025, the Authors. - https://creativecommons.org/licenses/by/4.0/
Download date	2026-11-01 22:12:46
Item downloaded from	https://hdl.handle.net/10468/19176

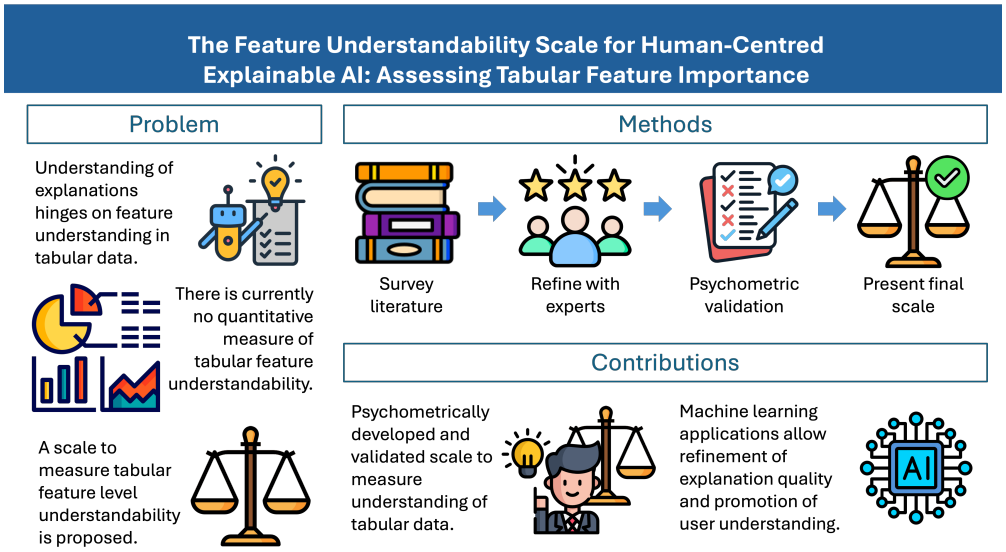
UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

arXiv:2510.07050v2 [cs.HC] 10 Oct 2025

Graphical Abstract

The Feature Understandability Scale for Human-Centred Explainable AI: Assessing Tabular Feature Importance

Nicola Rossberg, Bennett Kleinberg, Barry O'Sullivan, Luca Longo, Andrea Visentin



Highlights

The Feature Understandability Scale for Human-Centred Explainable AI: Assessing Tabular Feature Importance

Nicola Rossberg, Bennett Kleinberg, Barry O'Sullivan, Luca Longo, Andrea Visentin

- A validated scale assessing tabular feature understandability is presented.
- The scale quantifies the understandability of tabular features.
- The scale identifies understandable features, improving AI explanations.
- This permits the co-optimisation of explainability and understandability.

The Feature Understandability Scale for Human-Centred Explainable AI: Assessing Tabular Feature Importance

Nicola Rossberg[a,b], Bennett Kleinberg[c,d], Barry O'Sullivan[a,b,e], Luca Longo[a], Andrea Visentin[a,b,d]

[a] *School of Computer Science and Information Technology, University College Cork, Cork, Republic of Ireland*
[b] *Taighde Éireann - Research Ireland Center for Research Training in Artificial Intelligence, University College Cork, Cork, Republic of Ireland*
[c] *Department of Security and Crime Science, University College London, United Kingdom,*
[d] *Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands*
[e] *Insight RI Centre for Data Analytics, University College Cork, Ireland,*

Abstract

As artificial intelligence becomes increasingly pervasive and powerful, the ability to audit AI-based systems is growing in importance. However, explainability for artificial intelligence systems is not a one-size-fits-all solution; different target audiences have varying requirements and expectations for explanations. While various approaches to explainability have been proposed, most explainable artificial intelligence methods for tabular data focus on explaining the outputs of supervised machine learning models using the input features. However, a user's ability to understand an explanation depends on their understanding of such features. Therefore, it is in the best interest of the system designer to try to pre-select understandable features for producing a global explanation of an ML model. Unfortunately, no measure currently exists to assess the degree to which a user understands a given input feature. This work introduces two psychometrically validated scales that quantitatively seek to assess users' understanding of tabular input features for supervised classification problems. Specifically, these scales, one for numerical and one for categorical data, each with two factors and comprising 8 and 9 items, aim to assign a score to each input feature, effectively producing a rank, and allowing for

the quantification of feature prioritisation. A confirmatory factor analysis demonstrates a strong relationship between such items and a good fit of the two-factor structure for each scale. This research presents a novel method for assessing understanding and outlines potential applications in the domain of explainable artificial intelligence.

Keywords: Explainability, Interpretability, Psychometrics, Scale Development, Understandability

---

1. Introduction

In recent years, artificial intelligence (AI) has expanded its capabilities and become increasingly pervasive throughout society. With this increase in potential, a growing number of fields are exploring ways to harness the power of AI. However, this growing interest does not come without risk. Many Machine Learning (ML) models are black- or grey-boxes, where their size and complexity limit the degree to which humans can audit, analyse and understand decision pathways [19]. As such, ensuring consistent and theoretically sound inferences is becoming increasingly challenging. To provide insights into decision pathways, permit auditability of ML models, and prevent bias, a range of explainability methods have been proposed [59]. These methods promote both the legal compliance and the safe training of ML algorithms, especially in critical fields such as medicine [39].

While a wide array of eXplainable AI (xAI) techniques has been proposed, their adoption and type of explanation depend on the model type and data structure. One common type of explanation is textual, which is a natural language statement that expresses why a given output was produced by an ML model. These explanations are especially common for tabular data, where the output is explained through reference to the input features and their values [57]. For example, when applying for a loan, the input features may include 'income', 'debt', and 'character', among others. If the loan is rejected, the textual explanation may state that "The application was rejected because the applicant's income was too low". Here, the rejected loan is the output and "income" is the feature used to explain such a decision. In this explanation, a user's ability to understand why their loan was rejected hinges on their understanding of "income" as a key feature. As such, it can be extrapolated that the degree to which users can understand textual explanations for tabular data depends on their ability

to understand the features included in the explanations. For example, while "insufficient income" is a relatively straightforward feature to understand, a loan rejection based on 'low character' may be more difficult to grasp, and an explanation centred on this feature would not be helpful to the user. User understanding of explanations is vital, as research has shown that good explanations can promote both user trust and acceptance of the ML outcome [55]. To make models interpretable by design [11], it is hence important to favour features during model construction that are both relevant and understandable in explanations, ensuring clear communication with the target audience. However, identifying understandable features is challenging, as it is a subjective process that depends on the knowledge of the target audience, and currently, no tool exists that quantifies feature-level understandability.

This work fills this gap by developing and validating a questionnaire-style scale that measures the understandability of features in tabular data. The proposed *Feature Understandability Scale* is developed based on the literature and psychometrically validated using the ten steps for scale development outlined by Carpenter [10]. The goal of the scale is to measure the degree to which a specific user group can understand tabular features, thereby allowing explanations to be tailored to their level of understanding. To the authors' best knowledge, this is the first scale that permits the quantification of user understanding, assigning an 'understandability' score to each feature in a tabular dataset. This article resides within the first step in the ML workflow shown in Figure 1 and contributes to the literature on interpretability-by-design. To summarise, this research focuses on:

1. Conducting an extensive literature review on the existing taxonomies and needs of xAI stakeholders is conducted;
2. Designing a new scale measuring feature-level understandability is proposed and psychometrically validated;
3. Proposing future avenues for the integration of feature-understandability scores for interpretability-by-design solutions.

The design and validation of psychometrically sound measures hinge on proper scale development and reporting, which is frequently absent or inconsistent in the literature [10, 48]. The disparities resulting from the lack of standards lead to questionable reliability and validity, as well as significant quality discrepancies between developed scales. To ensure high

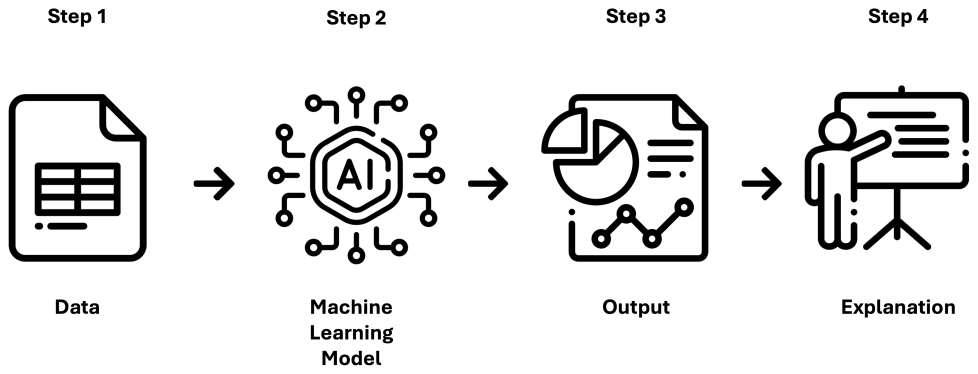


Figure 1: Overview of the xAI workflow for tabular data. The current work contributes to Step 1 by providing a way to measure 'understandability' for each feature.

standards during the development of this study, our methodology is based on Carpenter [10]'s ten steps for scale development, which are detailed in Section 3.

The remainder of this manuscript is structured as follows. Section 2 first defines the relevant concepts for this review, ensuring a consistent frame of reference for key terms. It then summarises the known findings on xAI stakeholder taxonomies and needs. Based on such a review, an initial scale is proposed. Section 3 describes the methodology for the scale refinement and subsequent psychometric validation following the ten steps proposed by Carpenter [10]. Section 4 presents the outcomes of the design and validation steps as well as the scale alterations resulting from each step. Finally, Section 5 analyses the most important findings of this study and presents avenues for applications and future research.

2. Literature Review

This section is divided into three key parts. First, in Section 2.1, the relevant definitions are extracted from the literature to ensure a consistent understanding of key terms going forward. Next, Section 2.2 reviews the existing stakeholder taxonomies and the known stakeholder needs in xAI. On this basis, the initial scale is designed and presented as *Draft 1* in the additional materials.

2.1. Definitions

The definition of explainability in the literature is broad and frequently contradictory [46, 39]. Some definitions identified based on conversations with stakeholders and the existing philosophical literature include "the ability to explain or present in human terms" [15, 46] and "to provide information about [an event's] causal history" [37]. In their systematic review, Muhammad and Bendechache [42] combine these definitions, arguing that explainability refers to "the extent to which humans can understand an AI model's decision-making process, providing clear and interpretable insights into how the model arrives at its conclusions or decision, facilitating user trust and validation of the results". Others focussed less on the definition of explainability but rather on the elements which should be included in a good explanation; Halpern and Pearl [21] argue that a good explanation provides an answer to a 'why' question, Hoffman et al. [26] emphasizes that a good explanation is required to be clear and precise, such that an a-prior judgement regarding the quality of an explanation can be made.

Explainability. Based on this brief survey of the literature, the following definition for explainability is extracted, which will be carried forward throughout the remainder of this manuscript: "Explainability is a model's ability to provide information about a causal relationship between the input and outcome, promoting the explainee's understanding". This research specifically focuses on the aspect of promoting understanding, as not all explanations linking input and output successfully promote user comprehension.

Understanding. In the literature, the concept of 'understanding' is more vague than that of explainability, despite definitions of explainability frequently hinging on the concept of understanding, with researchers defining explanations as 'statements that make AI systems understandable for users' [8, 2]. In this study, understanding is defined as the "ability to grasp a given concept to the extent that it may be adequately explained to another individual".

Measurement. Measurement is defined as the "quantification of the amount or extent of a given object or concept". Measurement is closely related to the concept of operationalisation, which involves converting an abstract construct into a measurable one by identifying a proxy measure. Going forward, abstract constructs will be referred to as concepts, while the operationalised measure will be referred to as a feature.

Outcome. One frequently referred-to concept is the 'outcome' of an ML model. This is especially relevant regarding participants' interest in accepting or changing given outcomes. In the context of this research, the outcome refers to the prediction or classification that an ML model has made during a supervised learning task based on tabular data. Referring to the loan application example, the outcome would be the granting or rejection of a loan.

2.2. Theoretical Concepts

To identify which aspects of explainability are considered important by different user groups, the literature was reviewed to identify relevant stakeholder groups and their respective needs. The following manuscripts focused on reviewing existing stakeholder taxonomies and subsequently analysing their xAI needs.

Stakeholder Taxonomies. One important consideration in measuring feature understandability is the variation based on domain and target group. The variety of users and corresponding needs in xAI has been repeatedly confirmed in the literature [35, 50]. Throughout these past works, researchers argue that, depending on the target audience of the proposed system, the chosen method of explainability should be adapted to optimise understanding in the user group. The following focuses on reviewing the different taxonomies of users stipulated in previous research and analysing the corresponding explainability needs. Figure 2 showcases how the different taxonomies overlap to create distinct stakeholder groups.

Preece et al. [45] stipulated that the four types of explanation-users should be broken down into developers, theorists, ethicists and end-users. The authors define developers as those individuals responsible for designing and deploying AI structures in the industry, who require explainability for quality assurance. Theorists are individuals who work on the understanding and advancement of AI theory, primarily located within the field of academia. The primary motivation for explainability in this group is the enhanced understanding of ML models. Ethicists are people concerned with the fairness of AI systems, and the majority of this group's members are from outside the field of computer science, including policymakers and governance entities. Members of this group seek explanations to ensure the fair and unbiased implementation of AI, thereby ensuring accountability and auditability of the AI-based system. The final user group comprises end-users of the system, who are individuals who may or

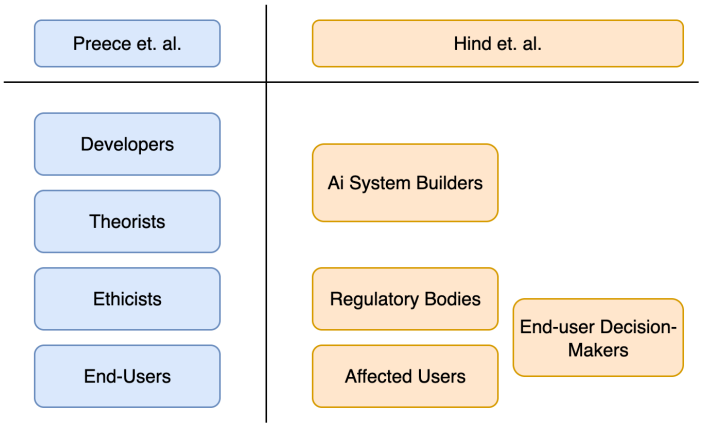


Figure 2: Overview of stakeholder taxonomies.

may not belong to the previous groups, but whose motivation for seeking explainability is to make and justify actions based on the prediction of an ML model.

A different taxonomy was proposed by Hind et al. [24], who hypothesised that users should be divided into end-user decision-makers, affected users, regulatory bodies, and AI system builders. The first group comprises individuals who use the output of AI systems to make impactful decisions, including physicians, judges, and law enforcement officers. These stakeholders require explanations to build faith in the system and potentially acquire additional insights to support future decisions. The second group of affected users are those impacted by the decisions of the previous group, including patients, loan applicants, and individuals who have been arrested. This group requires explanations to assess whether their treatment is fair and to identify the factors that may impact it. The third group of regulatory bodies comprises governing bodies which require explainability to ensure citizen safety. A prominent example of such a regulatory body is the European Union and the recent publication of the AI Act, which stipulates that explainability is legally required for any high-risk applications [17]. The final group of 'AI system builders' mirrors Preece et al. [45]'s group of developers and theorists, consisting of individuals in industry (or academia) who are responsible for deploying high-quality AI and require explainability to ensure the soundness of the developed system to promote understanding and advancement of AI theory.

The authors recognise that both taxonomies of explainability users have merit, identifying and dividing the most important user bases by explainability usage and requirements. However, what these differing approaches emphasise is that explainability needs differ greatly between user groups. As such, it is essential to tailor the methodology for producing explanations to the needs and level of understanding of the target group.

A wide array of xAI techniques tackling the problem of explainability in AI-based systems and varying explanation needs have been proposed. Although a comprehensive review of existing xAI techniques is beyond the scope of this research, several previous works have conducted systematic reviews on the topic of existing xAI approaches across various domains. Wang et al. [59] provides a comprehensive overview of the existing stakeholder groups, their respective explainability needs and the xAI methods suitable for meeting these needs. Saranya and Subhashini [49] review the current state of the art of existing xAI models and application domains. Regarding domain-specific methods, the most commonly analysed area is the medical field, with relevant reviews including Muhammad and Bendechache [42], Contreras and Bocklitz [12], Antoniadi et al. [5], Band et al. [6]. Other domains relevant for xAI approaches include finance, as reviewed by Weber et al. [60] and transportation, as analysed by Nwakanma et al. [44]. While a wide range of xAI methods exist across these domains, the majority use features to contextualise a model's output when working with tabular data. Consequently, the quality of the produced explanations depends on the user's understanding of the features of the input data. Consequently, to have successful explanations of feature importance, features must be chosen for audience understanding, taking into consideration that this selection will differ depending on the chosen target audience.

Stakeholder Needs. In addition to analysing the types of stakeholders, the literature on the needs of xAI stakeholders was reviewed. Based on the analysis of existing needs and requirements for explainability, three key conclusions were drawn.

The first key finding was that stakeholders are interested not only in understanding the model algorithms but also in the input data used to train the model [23, 25, 28, 34]. Two important aspects emerged under this heading. The first was the importance of understanding the variable within the context of the users' work, as well as the given outcome variable. The second was the understanding of the measurement of the feature

and how it was positioned with respect to the range of possible values. A brief review of the supporting research is provided below. In the work by Hepenstal and McNeish [23], 12 stakeholders were interviewed regarding their opinions on explainability in two hypothetical scenarios. The most common theme extracted from all interviews was a focus on understanding the input data and the quality of measurement. This emphasises the importance of not only understanding the individual features but also the method of measurement and the degree to which these measures are valid and reliable. In a similar strand of research, Hoffman et al. [25] interviewed 18 individuals who are members of varying stakeholder groups of xAI to gauge their opinions on the explanations and understanding of AI. Based on the structured interview process, the authors concluded that stakeholders expressed greater interest in knowing and understanding the input data of a model than in the inner workings of the ML models. The research by Kim et al. [34] found that both practitioners and end-users are interested not only in the features relevant to a given prediction but also in the 'reference range or values' for these features. This emphasises that users need to understand which features contributed to a given output, as well as be able to contextualise these features' values. While this may be intuitively possible for some features, others can be more difficult to understand, emphasising the importance of identifying and prioritising explainable features. For example, while 'income' may have intuitively understandable ranges, values of 'character' may be more difficult to grasp. Finally, Hong et al. [28] found that system developers are already considering interpretability at the design stage, taking into account how certain feature engineering choices will affect not only model performance but also interpretability. These past works highlight that stakeholders value understanding of the training data equally or more than understanding of model structure, as it provides a more contextualised insight into model performance.

A second important finding of the literature review is the significance of tailoring explanations to the educational and vocational background of the chosen target audience. The process of explaining is considered a knowledge transfer process, and consequently, explanations need to be adapted to the knowledge base and context of the user [16]. Hind et al. [24] proposed that a good explanation should fulfil three criteria. It should justify a decision to increase trust in said decision, the complexity of the explanation should match the 'complexity capacity' of the target audience, and the

explanation should match the domain, incorporating the domain-relevant terms. Given the context of the current research, the second and third requirements are applicable, emphasising that any provided explanation should be tailored to both the understanding and context of the explainee. As explanations are often reliant on feature importance, the employed features must satisfy these criteria [7]. Moreover, during the interviews conducted by Hoffman et al. [25], stakeholders overwhelmingly emphasised the importance of tailoring explanations to the audience's knowledge base. Not only did interviewees consider it important to take into consideration the background and understanding of users, but they also highlighted the importance of making explanations self-contained, with one stakeholder stating "Can a user, without training, figure out how to use the system within 10 min? If they fail at that, they don't use it". Based on this, it is clear that explanations must take the user's understanding into account. These studies emphasise that audience-tailored explainability is not only theoretically important but also satisfies a real-world need that has been repeatedly expressed by stakeholders from diverse backgrounds.

The last conclusion drawn based on the literature is that stakeholders hold widely divergent conceptualisations of explainability, and explanations should be adjusted to the expectations of the recipients to improve understanding. Brennen [8] interviewed 60 stakeholders across various domains regarding their understanding of and demands for xAI. Interviewees included developers in academia and industry, policymakers, investors and end-users. One of the key findings of the research was that there was no consistent understanding of explainability among groups, with stakeholders using terms such as 'verifiable', 'unbiased', 'self-service', and 'repeatable' when referring to xAI. The author argues that this emphasises a lack of consistent terminology, which may hinder discussion. While the discrepancy in users' understanding of explainability is certainly an artefact of miscommunication about the intent and ability of explainability, it also provides insight into what stakeholders desire to gain from explanations. Based on the synonyms provided, it is concluded that users wish to be able to understand and verify measurements independently ('self-service', 'verifiable', 'reliable'), comprehend the meaning of explanations without support ('understandable', 'ready-to-use'), and understand model decisions and feature usage ('auditable', 'accountable'). This emphasises the importance of tailoring explanations to the outcome group, as different stakeholders hold varying expectations of system explanations. Lopes

et al. [40] developed a taxonomy for evaluating the methods used to assess xAI systems, based on prior research and stakeholder needs identified in the literature. The authors highlight four key concepts relevant to human-centred explainability: trust, explanation usefulness and satisfaction, understandability and performance. In the context of this research, this emphasises the importance of establishing user understanding of explanation components as well as encouraging the establishment of trust through confirmation of feature measurement and fidelity.

In addition to analysing the needs and desires of stakeholders, previous studies evaluating the quality of explanations were consulted during the development of the initial scale. The work by Holzinger et al. [27] proposes the system causability scale, which provides a quick and simple method of assessing the causability of an explanation interface or an explanation process in itself. The system causability scale was based on the system usability scale originally proposed by Brooke [9], which was designed to measure the usability of an industrial system in a 'quick and dirty' way. As such, it is essential to note that these scales were designed with the objective of brevity, rather than absolute validity, which reduces their applicability in the current context. In the same vein of research, Vilone and Longo [58] proposed a 12-item questionnaire to evaluate the quality of explanations produced by xAI approaches. This research marks an important step in the validation of explanations, as it utilised psychometric methods to ensure the quality of the designed scale.

3. Methods

After the initial design of the scale based on the retrieved literature, several steps are taken to refine the structure and item phrasing before validation, following the methods by Carpenter [10]. However, it is important to note that this study differs from traditional scale validation due to the nature of understanding. Where the validation of scales in traditional psychometrics depends on variations in participants alone, such as individual changes in narcissism, depression, the current study considers variations in both participants and features, as understanding hinges on their interaction. As such, the sample size and the resulting analysis were adjusted to consider not only the item-to-participant ratio but the item-to-feature-to-participant ratio, marking a novel contribution to psychometric development.

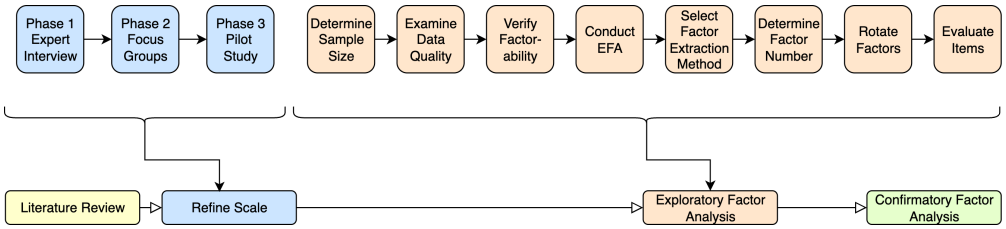


Figure 3: Workflow of the Psychometric Development of the Scale

Going forward, expert interviews were conducted to assess their understanding of the concept of explainability and to review the initial scale, identifying missing and redundant items. Next, two layperson focus groups were conducted to refine the phrasing and clarity of the items. Finally, a pilot study involving 89 participants was conducted to gather broader, anonymous feedback on the structure and clarity of the scale and to make any necessary final amendments before commencing large-scale data collection. Ethics approval for all studies was granted by the University College Cork Social Research Ethics Committee under log numbers 2024-178 and 2024-166A1 and complied with relevant laws and institutional guidelines. Below, each step is described in further detail. The complete workflow is illustrated in Figure 3.

3.1. Initial Scale

The initial draft of the scale was created based on the literature reviewed in Section 2. While the included studies employ a post-hoc approach to measuring explanation quality, their goal aligns with that of the current study, and their items were adapted for the initial draft of the *Feature Understandability Scale*. The ten categories proposed by Holzinger et al. [27] were evaluated, and those relevant to *Feature Understandability Scale* were identified. Out of the ten factors proposed by the authors, five were adapted to the feature-level domain and included in the initial draft of the scale. An additional five items were adapted from the questionnaire by Vilone and Longo [58]. All items were selected based on qualitative inspection and evaluation of their suitability for ad-hoc usage. In addition, the bank of questions that different stakeholders may pose to an AI-based system, as designed by Liao et al. [38], was consulted, and several items were adapted for the initial draft of the *Feature Understandability Scale*. The initial draft of the scale, based on the literature, can be found in the

additional materials (*Draft 1*). The following describes each step taken to improve, verify, and test the drafted scale.

3.2. Refining Scale

Phase 1: Expert Interviews. Expert interviews aim to gather feedback on the quality of items and the extent to which they accurately reflect the overarching construct. Interviews should start by probing broad definitions of the constructs before inquiring about their interpretation of each subscale and item of the scale. This feedback can be acquired through open-ended feedback or Likert-style ratings. To gather expert opinions, semi-structured interviews were conducted with six experts in the field. Subjects were sampled from industry and academia and held degrees in computer science, psychometrics, or related fields at the Master's level or higher. All subjects were working on explainability, human-machine interaction or related topics at the time of the interview. Subjects were identified based on the researchers' personal network and their academic credentials. Informed consent was obtained from all participants, and subjects were debriefed about the study's aim prior to the interview.

The interview consisted of two steps. First, participants were asked to respond to several questions regarding their general understanding of explainability, both at an item and explanation level. This aimed to supplement the understanding and definitions of stakeholders previously extracted from the literature. The second step involved presenting each item to the experts. The items were grouped and presented by subscale. Experts were first asked for their opinions on the subscale structure, whether any items should be added or removed from the scale, and the reasons behind their suggestions. Subsequently, they were asked to review the phrasing of each item to clarify the structure and improve the language quality. Based on this, the scale was refined, and the revised scale was presented in the next step.

Phase 2: Focus Groups. In addition to the feedback received during expert interviews, two focus groups were conducted, each with 2 and 3 laypeople, respectively. Based on the recommendations of Carpenter [10], this pre-test assessed the clarity of the phrasing of items, focusing especially on those that were ambiguous, leading, or confusing. The focus group structure was similar to that of the expert interviews, where participants were first asked to share their general understanding of xAI before being debriefed about the research goal and presented with the designed scale.

Based on the focus groups, item phrasings were adapted, and items were edited. The resulting scale was presented in Phase 3.

Phase 3 - Pilot Study. The final step of the initial scale design involved a pilot study implemented using Qualtrics (https://www.qualtrics.com/). This aimed to verify item structure, phrasing, and understanding through a larger, anonymous cohort. After completing the informed consent form, a brief explanation of the study's aim was provided, along with examples of features (both numerical and categorical), values, measurement scales, and endpoints for reference. Participants were then presented with each scale item and asked to provide any feedback, questions, or comments on clarity. Participants were recruited through convenience and snowball sampling, and no prior expertise in xAI or related fields was required, as feedback was sought only on item wording and structure.

3.3. Exploratory Factor Analysis

Exploratory Factor Analysis (EFA) is a statistical technique used to group data according to underlying latent variables [10]. It rests on the assumption that the observed data can be divided into several related groups and by analysing inter-correlations in items, these groups or 'factors' can be identified. It is frequently used in psychometrics to identify the latent constructs constituting a given measurement, as for instance grandiose and vulnerable narcissism functioning as factors of overall narcissism [51]). EFA is used to analyse and validate the constructed scale. The following describes the methodology employed in this study using EFA.

Determine Sample Size. Determining the correct sample size is crucial to ensure the study has sufficient statistical power. Power refers to the ability of a study to reject a false null hypothesis while maintaining sufficient confidence in the result [4]. While Worthington and Whittaker [62] argues that a minimum sample size of 300 may generally suffice, other research suggests that the sample size should be determined by the item-to-participant ratio rather than absolute values. Early recommendations by Gorsuch [18] proposed ratios of 1:5 or 1:10 items to participants. However, more recent research by Costello and Osborne [13] increases these recommendations to 1:20 as this produced more robust and correct solutions in their research [48].

These recommendations served as a basis for determining the sample size in the present study. However, this work differs from existing scale-construction articles in that the concept of understanding does not solely rest on the respondents but also on the features to which the scale is applied. As such, it is important to not only consider the ratio of items to participants, but also that of items to features. After consulting the literature and discussing with several domain experts, it was decided to select 9 features across three relevant xAI domains (medicine, loan and hiring decisions) and evaluate them using the developed scale. For each feature, 240 ratings were collected from participants, resulting in a total of 2160 ratings across the 9 features. After splitting the sample into EFA and CFA, this corresponds to a 1:47 (numerical) or 1:55 (categorical) item-to-response ratio. This exceeds existing recommendations and ensures effective validation.

We used a large-scale general population survey to collect data for the scale development. Participants were recruited via Prolific[1]. Participants residing in the United States, the United Kingdom, Ireland, Canada, Australia and New Zealand were recruited. Each participant completed an informed consent form during which they were debriefed regarding the purpose of the study. Subsequently, participants were asked to complete the questionnaire. Each participant was assigned three of the nine features and asked to rate them using the developed scale. Features were presented pseudo-randomly and with equal frequency, ensuring that the same number of ratings were collected for all features. All items were measured on a 5-point Likert Scale, ranging from 'strongly disagree' (1) to 'strongly agree' (5). Features were selected from xAI-relevant areas as identified by Islam et al. [30] and taken from domain-specific xAI research [33, 41, 36, 63, 43, 20]. Table 1 shows the selected features and their intended meaning.

Examine Data Quality. To ensure the reliability of our findings, several pre-checks are implemented to test data quality before analysis. These checks ensure that low-quality or non-human responses are identified and excluded. Attention checks were implemented for each feature evaluation, asking participants to select the 'neutral' option for one item. If individu-

[1]Prolific connects researchers with participants, enabling the distribution of surveys and participant reimbursement. Link: https://www.prolific.com/

Table 1: Features Selected for Scale Testing

Domain	Feature	Intended Meaning	Numerical or Categorical	Source
Hiring	Recruitment Strategy	Strategy adopted by the hiring team for recruitment	Categorical	Kharoua [33]
Decisions	Personality Score	Score of candidate's personality traits	Numerical	Kharoua [33]
	Degree	Highest educational degree held	Categorical	Mahmoud et al. [41]
Medicine	BMI	Body Mass Index	Numerical	Lee et al. [36]
	Total Cholesterol	Cholesterol in mg/dL or mmol/L	Numerical	Lee et al. [36]
	Neoplasm Stage	Whether the tumour is regional or has spread	Categorical	Namdari [43]
Loan	Credit Score	Applicant's credit score	Numerical	Xu et al. [63]
Applications	Debt	Total amount of individual's debt	Numerical	Xu et al. [63]
	Business or commercial	Type of loan applied for	Categorical	H. [20]

als fail the attention check, their response for this feature is discarded. The following metrics were consulted to assess data quality before analysis:

1. Attention checks [1];
2. Completion time (very fast time may indicate lack of focus) [48];
3. Missing Responses [10].

Data preprocessing was conducted in Python, and the analysis was conducted in R using the 'psych' package [61]. The complete code used for analysis can be found on Github[2].

Verify the Factorability of the Data. The next step in EFA is to check whether the data can be factored. Factorisation refers to the identification of underlying subscales that group the items. To verify that data can be factored, the following tests were inspected:

1. Correlation Matrix - values of 0.30 or higher [10]
2. Bartlett's Chi-Square test with $p \leq 0.05$ [52]
3. Kaiser-Meyer-Olkin (KMO) of 0.60 or higher [52]

Based on these tests, violating items were inspected, and deletion was considered if the items were inconsistent with the overarching concept.

Conduct Exploratory Factor Analysis. After cleaning and confirming factorability, EFA was conducted. It identifies factors of common variance in the data. The following steps were taken to run EFA for the construction and validation of the final *Feature Understandability Scale*.

[2] github.com/ncrossberg/Feature-Understandability-Scale

Select Factor Extraction Method. Factor extraction identifies latent variables. There is limited information regarding the benefits of different extraction methods. Since the chosen sample size in this study exceeds recommendations, 'maximum likelihood' is selected for factor extraction.

Determine Number of Factors. After factor extraction, the number of factors to retain was determined. Generally, it is recommended to either choose one method in advance of data analysis or examine convergence between different methods. Here, the convergence between an 'eigenvalue greater than 1', and the 'scree elbow test' is examined to determine the number of retained factors [10].

Rotate Factors. Rotation transforms the factor loading matrix to simplify its structure and ease the identification of item-factor relationships. Rotation can be conducted in either an oblique or an orthogonal fashion. Orthogonal rotation produces uncorrelated factors, which is unexpected in the present case, as factors are intended to measure understanding and are therefore likely to be correlated. Oblique rotation is hence implemented in this study. There are several types of oblique rotation, including 'Direct Oblimin' and 'Promax'. 'Promax' is chosen, as it begins with an orthogonal solution and then transforms it to an oblique solution, making it more robust [56].

Evaluate Items Based on A Priori Criteria. After these analyses, the number of items should be reduced. Ideally, the initial scale should include 3 times as many items as necessary. The following criteria were used to reduce items:

1. Minimum factor item loading > 0.32 [10];
2. Items cross-loading > 0.32 on other factors are excluded [13, 53];
3. Minimum number of 3 items per factor [53];
4. Examination of theoretical convergence of items [10];
5. Parsimony to minimise the redundancy of wording or meaning across items [10].

3.4. Confirmatory Factor Analysis

In addition to the 10 steps proposed by Carpenter [10], Confirmatory Factor Analysis (CFA) ensures that the scale developed during EFA can be identified in new data, further validating its structure. The methodology for CFA in this study is based on Rottweiler and Gill [48].

Participants. Data collection for CFA was conducted simultaneously with EFA, and the data were split evenly between the two procedures. For each feature, 240 ratings were collected, resulting in a total of 2,160 ratings for CFA. The methodology was the same as that of EFA, with respondents being recruited via Prolific, completing an informed consent form and being asked to rate three features using the developed scale.

Procedure. CFA aims to verify whether the factor structure identified during EFA remains consistent in new data. The data was cleaned using Python and then analysed in R using the packages 'Lavaan' and 'SemTools' [47, 31]. After completing the analysis, the fit indices were examined and evaluated against the following standards put forward by Hu and Bentler [29]:

1. Chi-Squared (χ^2) significance with $p < 0.05$;
2. Comparative Fit Index (CFI) ≥ 0.9;
3. Tucker Lewis Index (TLI) ≥ 0.9;
4. Root Mean Square Error of Approximation (RMSEA) $\leq .08$;
5. Standardised Root Mean Square Residual (SRMR) ≤ 0.08.

The data displayed a skewed distribution, and a robust estimation using 'Satorra Bentler' was consequently implemented. 'Satorra Bentler' is a chi-squared estimator, which accounts for non-normality and corrects for heteroscedasticity of data.

4. Results

4.1. Refine Scale: Expert Interviews

The experts were posed several questions regarding their understanding of explainability, what they wish to gain from explainability, and their opinions on user needs for explanations. The full script for the semi-structured interviews is available in the additional materials. The primary goal of the interviews was to gain a deeper understanding of experts' opinions and perspectives on explainability, and to solicit their thoughts on the developed scale. Three core topics were extracted from the responses, and new items were synthesised on their basis. Each topic is detailed below, and examples of the extracted items are provided. It is important to note that several similarly phrased items are included, as this allows the better-functioning item to be selected during quantitative analysis. The changes made to the scale based on the expert interviews are summarised in Table 2.

Inputs. The most frequently mentioned factor related to the quality and understanding of explanations was the reference to inputs and the data used. Generally, experts remarked that xAI 'allows challenging decisions based on features' and shows 'which features an algorithm is using for predictions'. One important point not previously identified in the literature was that the use of only the overarching features may not always be sufficient. Rather, the value of the feature should be included in the explanation to provide nuance regarding the feature's influence. For instance, returning to the loan example discussed earlier, an applicant may be interested in *how much* their income is, rather than that it is simply too low. This would further enable the user to understand their position in relation to the possible range of feature values and relevant turning points. This highlights the importance of users being able to understand the scale of features included in the explanation, as well as the endpoints of the measurement and how their data is positioned on this scale. To capture the importance of various values of features, the 'understanding' subscale is expanded with items referring to both the user's understanding of their value on the feature and the range of the feature, as well as the direction of influence of several values.

Another pertinent insight was that features do not function in isolation but rather should be considered in their interaction with other features in the dataset, and how this impacts the algorithm's performance. As such, users should at the very least be aware that features may interact with each other and how this interaction may influence the outcome. While this was, to a certain extent, already captured in the faithfulness subscale, several additional items were constructed to test alternative phrasing and capture additional facets of item interaction.

A final consideration was the degree to which features are 'actionable' if the users wish to overturn the given decision. 'Actionability' refers to the degree to which the user may change their value on a given feature. While explanations may be faithful to the system, they may not be helpful to the user, as impactful features cannot be changed [58]. Hence, it may be important to measure the user's perceived level of control over the value of the feature, and the scale is adapted accordingly.

Outputs. A second, frequently mentioned aspect was the output produced by the xAI system. The experts were queried regarding their expectations of explanations and what they think contributes to the understanding of an explanation. One important aspect was that explainability should not

only refer to the explanation itself but rather to the 'capacity of ML to motivate or justify decisions', indicating that several explanations should be producible depending on the context and question of the user. Closely connected to this was the algorithm's ability to answer follow-up questions, further refining the initially produced explanation. Furthermore, experts emphasised that they would like to see what possible outcomes may exist, as well as which datapoints may produce outcomes similar to those of the given user. Additionally, one interviewee highlighted that they would like to be informed of the algorithm's performance as part of the explanation to gauge the level of trust the outcome may warrant.

Communication. The final aspect that stood out across several interviews was the importance of effective communication in the explanation. Interviewees repeatedly emphasised that xAI should aim to foster users' trust, both in the algorithm and the explanation. To facilitate this trust, explanations need to be both clear and match the level of expertise and communication of the explainee. As such, the explanation's level of detail and structure must be adapted to its intended user. Examples of considerations given by the interviewees included details regarding the ML system, which information is extracted from the data, and the features included in the explanation. Based on this, it is essential to measure both the user's trust in the feature and its effectiveness, as well as the user's interest in the feature's impact on the outcome, as this may indicate the level of detail at which the feature functions.

After the general explainability questions, experts were presented with the developed scale and asked to review the subscales and items. The majority of feedback pertained to rephrasing certain aspects of the scale for clarity. One expert pointed out the circularity of questions U1 and U2, using 'understanding' to assess the 'Understanding' subscale, and suggested that these items be rephrased or deconstructed to present different aspects of understanding. They furthermore emphasised that questions employing two verbs should be disentangled to clarify which concept they are querying.

One item considered highly relevant but ambiguous by several experts was F2; "To my best knowledge, the feature is the main cause of the decision". This item was aimed to measure whether the user believed the feature to be the main determinant of the outcome, capable of 'overwriting' other contradicting features. Referring back to the loan example, one instance of this may be that a sufficient income can override other fac-

tors, such as a young age or poor credit score, during a loan application. This item is rephrased to "To my best knowledge, the feature affects the outcome more than other features".

Table 2: Scale additions based on the expert interviews.

Subscale	Source	Item Added
Understanding	Inputs	I know which values a feature can take.
		I know the endpoints of the scale.
		I know the meaning of the endpoints of the scale.
		I know how a change in value would impact the outcome.
Measurement	Communication	I trust that the feature is measured correctly.
	Scale Alterations	I require no support to understand the measuring scale of the feature.
		I require no references to understand the measuring scale of the feature.
Faithfulness	Inputs	I know how this feature may influence other features.
		I know how this feature may affect other features.
		I know how this feature may be combined with other features.
		I could change the value of this feature with reasonable effort.
		I think changing the value of this feature would change the outcome.
		I think I can change the value of this feature in a way that changes the outcome.
		I think a change in the value of this feature should change the outcome.
	Communication	I trust that the feature is important for the outcome.
		I am interested in how this feature affects the outcome.

The experts also recommended adding definitions at the top of the scale to clarify certain words, including 'feature', 'value', 'outcome' and 'scale'. These definitions will be provided to users before scale completion to ensure consistent understanding between users and minimise discrepancies caused by subjective interpretation of phrases. The revised scale can be found in *Draft 2* in the additional materials.

4.2. Refine Scale: Focus Groups

The first concept which arose from the discussion about general explainability was that of bias and fairness. One participant stated that subjects should rate the extent to which they believe the feature to be a fair way of measuring the underlying concept and whether the feature may introduce bias into the system. The argument behind this idea was that subjects should be able to attest to problematic features, such as postcodes for mortgage predictions, and that this should be taken into

consideration during system design. However, while this is a valid consideration, it is not related to the subjects' understanding of the feature and, as such, is not included in the current scale.

The second consideration was drawn from the research area of counterfactuals, aiming to probe the participants' understanding of how different values of the feature may impact the outcome. It was argued that, as a component of understanding, participants should not only be aware of which features impacted the outcome but also grasp how changes in the values of a given feature may change or affect the outcome. An example of this is that participants not only understand that debt affects the likelihood of being granted a mortgage, but also that a larger amount of debt decreases the likelihood, and vice versa. This is introduced into the scale under the 'Understanding' subscale using the following two items. Note that the alternative phrasings are introduced to allow analysis of item functioning, and only one phrasing will be maintained in the final scale.

1. I know how a change in the feature's value would impact the outcome.
2. I know how a change in the feature's value would change the outcome

The concept of 'Others'. The most prominent issue identified during the focus groups was the concept of 'Others' included in items such as 'I expect others affected by this decision to easily understand this feature'. This aimed to denote other individuals affected by the given decision, for example, other individuals applying for a loan, regardless of their respective outcomes. This was initially found to be challenging during the expert interviews. However, after clarification, a consensus was reached to retain the relevant questions. Additionally, during the focus groups, several participants noted that they were not in a position to speculate about the understanding of other individuals who were impacted by the decision. As such, items including this concept will be removed from the scale, as it has been identified as problematic during all stages of refinement.

The concept of 'Trust'. A second cause of confusion during the focus group was the use of the word 'trust' throughout the scale in items such as 'I trust that the features are measured using an objective scale'. Participants found the word to be both subjective and ambiguous, arguing that it should be replaced with phrasing akin to 'I think' or 'I assume'. It was

argued that the use of 'trust' is circular, as explanations should inspire trust, rather than being based on pre-existing trust in the system. As such, these items were rephrased accordingly.

The concept of 'References'. Several items in the scale refer to 'references' as a way of clarifying given features or concepts. This was a point of contention during the focus groups, with participants struggling to understand who or what the word 'references' referred to. This concept was previously flagged as problematic during several of the expert interviews. Consequently, relevant items are removed from the scale going forward.

The concept of 'Expertise'. Another point of clarification raised by participants was the question of what constitutes 'expertise'. One example of this is the item 'I think that this feature can be easily understood by an expert in the field' in the 'Understanding' subscale. Participants argued that both expertise and the field in question are subjective concepts, which are likely to introduce bias into the measurement. As it is difficult to provide definitions that will be universally applicable, the pertinent items are removed from the scale.

The concept of 'Reasonable'. The final concept that lacked clarity was that of 'reasonability' in items such as 'I think I can verify the value of the feature with reasonable effort'. The phrasing of 'reasonable' was perceived as highly subjective by all participants, and it was proposed that this be replaced with a more specific phrasing, such as 'within five minutes' or 'through an online search'. However, one issue with this is that the replacement of 'reasonable' would depend on the subjective interpretation of the focus group and authors of what may be considered reasonable. As such, it was decided to replace the phrasing with the concept of 'feasibility' and rephrase the items as follows: 'I think it is feasible for me to verify the value of the feature'.

In addition to the five overarching changes outlined above, some minor issues were identified during the interviews, which led to minor adjustments in the item order and phrasing throughout the scale. Several items that were perceived to measure the same construct were removed, retaining only the clearer formulation. The final adjusted scale can be found in the additional Materials labelled *Draft 3*. In addition to the changes in phrasing, several definitions were added to the scale to clarify concepts and ensure consistent understanding among participants. After

implementing the focus group feedback and revising each item individually for parsimony, 32 items remained and are presented as *Draft 3* in the additional materials.

4.3. Refine Scale: Pilot Study

The final step in constructing the scale was a pilot study to gather comprehensive feedback on the phrasing and structure of each item. In the pilot study, participants were asked whether each item was understandable and if they had suggestions for rephrasing or clarification. A total of 89 responses from lay individuals recruited via convenience sampling were collected and qualitatively analysed. Based on the responses, several items were rephrased for clarity, and three items were removed. The most common concern was item length and vocabulary complexity, and simpler phrasings and more accessible synonyms were utilised in the revised scale. After these revisions, the final scale consisted of 28 items and is presented as *Draft 4* in the additional materials.

4.4. Exploratory Factor Analysis

Data was collected on August 28th 2025, using the platform 'Prolific' in conjunction with 'Qualtrics'. The sample consisted of adults over the age of 18 residing in the United Kingdom, the United States, Ireland, Canada, Australia or New Zealand. A total of 1440 participants were recruited. Each participant rated three of the nine features of the scale shown in *Draft 4* of the scales found in the additional material, and ratings were randomly assigned and stratified. For each of the nine features, 480 responses were received (240 for EFA and 240 for CFA). The sample was split equally, with 720 participants used for EFA and 720 retained for CFA, stratified by the type of feature (numerical or categorical) to ensure equal distribution between EFA and CFA. This resulted in a total of 2160 ratings for EFA and CFA, respectively.

Examine Data Quality. Data quality was examined based on built-in attention checks, completion time, and missing responses. Attention checks were conducted at the feature level, and 19 responses were removed due to failed attention checks. The mean response time was 9.5 minutes (SD = 5.2). Response time patterns were examined using a box-and-whiskers plot, and no responses were deleted based on response time. There were

no missing responses. After preprocessing, a total of 2141 responses remained, comprising 1199 ratings for the numerical features (1:55 item-to-response ratio) and 942 for the categorical features (1:47 item-to-response ratio), which is sufficient for analysis.

Verify the Factorability of the Data. The factorability of the data was verified using Bartlett's test of sphericity, KMO assessment and correlation analysis. As the scale will vary depending on whether it is deployed for numerical or categorical features, these analyses were conducted separately, and two final scales are presented.

For the initial numerical scale, the Barlett's Chi-Square $\chi^2(231) = 16788.03$, $p \leq 0.001$, and the KMO $= 0.94$, and for the categorical scale, the Barlett's Chi-Square test $\chi^2(190) = 13268.69$, $p \leq 0.001$, and the KMO $= 0.97$. This indicates excellent common variance and multivariate normality in the joint distribution of the items [32]. Item-item, item-total and corrected item-total Pearson correlations were computed to assess whether items measured the same construct. The full correlation matrices are available in Appendix A. For the numerical scale, one item correlated at less than 0.3. Carpenter [10] recommends deleting items that correlate below 0.30 if it makes theoretical sense to do so. Since item 5 was theoretically consistent with the remaining items on the scale, it was retained at this stage. For the categorical scale, no item correlations fell below the 0.30 threshold in the item-total correlations, but two items potentially measure different concepts than the remaining items. All items were retained for the initial EFA. These results indicated that factoring is possible, and analysis moved forward to the initial EFA.

Conduct EFA. The initial factor analysis was conducted with Maximum likelihood for factor extraction and 'Promax' rotation. Based on examination of the scree elbow plot and eigenvalues, three factors were retained for the numerical scale and two for the categorical scale. The initial factor assignments can be seen in Figure A.5 in the Appendix. Factor loadings were examined to assign items to the retained factors, and item-factor correlations were recomputed.

Evaluate Items Based on A Priori Criteria. After computing the initial sub-scale structure of the two scales, the included items were evaluated based on a priori criteria. Items were removed if they had an item-factor loading < 0.32 (this removed two items from the numerical scale and none

from the categorical scale). If items cross-loaded > 0.32 on two factors, they were removed, leading to the deletion of one item from the numerical scale and two from the categorical scale. Factors with fewer than three items were then discarded. Finally, based on a qualitative examination of the theoretical convergence of items, as well as the parsimony in wording and meaning across items, subscales were reduced. The final scales had 2 factors each, with 8 (numerical) and 9 items (categorical), respectively.

Present Results. After item-reduction, the EFA was re-run using maximum likelihood factoring and 'Promax' rotation. The analysis revealed a two-factor solution for both scales, comprising 8 items for the numerical scale and 9 items for the categorical scale. No cross-loadings > 0.32 or weak loadings < 0.32 remained for either scale. For the numerical scale, the first factor consists of 5 items and the second of 3. For the categorical scale, the first factor consisted of 6 items and the second of 3 items. The sum of squared loadings refers to the factor's variances after their extraction [54]. For the numerical scale, the sum of squared loadings was 2.86 (28.6%) and 2.27 (22.7%), explaining a total of 51.3% of the total variance. The sum of the squared loadings for the categorical scale was 3.74 and 1.97 for each factor, respectively. This represents 37.4% and 19.7% of variance, explaining 57.1% of the total variance.

To assess scale homogeneity and ensure consistency between items measuring the same construct, corrected item-total correlations were recomputed [64]. In the literature, a value of ≥ 0.30 for each item-total correlation is considered sufficient; however, values between 0.30 and 0.70 are recommended. [14]. For the numerical scale, corrected item-total correlations ranged between 0.64 - 0.73 and for the categorical scale between 0.64 - 0.81. This indicates that both scales have good homogeneity and inter-item consistency. Additionally, communalities for all items were assessed. Communalities indicate the proportion of variance of each item, which is explained by the factors. Satisfactory communalities range from 0.40 to 0.70, and are considered good when above 0.7. For the numerical scale, communalities ranged from 0.46 to 0.76 with a mean of 0.64. For the categorical scale, communalities ranged from 0.58 to 0.74 with a mean of 0.65. As such, both scales met or exceeded the requirements for the proportion of item variance explained by the factors.

Finally, to ensure that no redundancy remained in the retained items, inter-item correlations were assessed [48]. The full item-item correlation tables are available in Appendix A. Ideally, correlations should range be-

Table 3: Metrics for the final scale after EFA.

	Factors	Mc Donald's ω	F1	F2
Numerical	F1	0.87	1	
	F2	0.90	0.48	1
Categorical	F1	0.92	1	
	F2	0.85	0.66	1

tween 0.20 and 0.50; items with correlations below 0.20 and above 0.80 should be excluded. The item correlations ranged from 0.24 to 0.75 for the numerical scale and from 0.35 to 0.72 for the categorical scale. The average item-item correlation was 0.52 for the numerical scale and 0.6 for the categorical scale. To assess the internal consistency of the scales, McDonald's omega was computed, which is considered more robust than Cohen's alpha [22]. The results, alongside the correlations of the factors, are presented in Table 3. Both scales demonstrated very good internal consistency and moderate to strong correlations among the factors. The final scales, alongside factor loadings and corrected item-total correlations, are presented in Tables 4 and 5. It is interesting to note that for both scales, Factor 2, 'Feature-Outcome Relation', consists of the same items. Only factor 1 varies as a function of the differences in variable conceptualisation (numerical or categorical).

The Feature Understandability Scale is split into two scales, which measure the understanding of numerical items and categorical items, respectively. Both scales are two-factor scales. For the numerical scale, the first factor consists of 5 items and the second of 3. The categorical scale's first factor comprises 6 items, and the second comprises 3 items. The factors for both scales are named under the same convention due to the similar grouping of items. Names were chosen based on an inspection of the included items and the nomenclature of the initial scale. Factor 1, 'Understanding and Measurement', contains items which measure the individual's ability to understand the feature itself as well as the method through which it was assessed. Factor 2, 'Feature-Outcome Relation', is composed of items assessing how the respondent views the relationship between the feature and the outcome variable. Within each scale, the factors showed a moderate to strong positive correlation [3]. The factors for

the numerical scale were correlated with $r = 0.48$, and for the categorical scale with $r = 0.65$.

4.5. Confirmatory Factor Analysis

The CFA aimed to confirm the factor structure identified during the EFA. The following statistics were tested: The chi-squared statistic was assessed, the CFI ≥ 0.90, TLI ≥ 0.90, RMSEA $\leq .08$, SRMR $\leq .08$ [29].

Data Cleaning. Before running the CFA, the data was cleaned. The procedure described in Section 4.4 was followed, using the second part of the data (n = 720). Responders missed attention checks for a total of 11 items, for which the responses were set to N/A. This resulted in a total of 2,157 ratings (1,198 for the numerical scale and 951 for the categorical scale). No answers were discarded due to slow response time or incomplete answers.

Results CFA. The CFA was run with a robust estimator, as both the numerical and categorical data showed skewness. The results of the confirmatory factor analysis are displayed in Table 6. Overall, the numerical model's fit is found to be acceptable, and the categorical model demonstrates excellent fit. The factor loadings for the scales ranged from 0.84 to 1.0 for the numerical scale and 0.79 to 1.0 for the categorical scale, demonstrating strong relationships. These results indicate a good fit of the two-factor scales, and the final CFA models are shown in Figure 4.

Based on the methodology in Rottweiler and Gill [48], the current 2-factor scales were compared to a one-factor solution, where all items loaded on a single factor. This was done to ensure that the chosen two-factor structure was the best-fitting model. For both scales, the one-factor fit was extremely poor, and the results are shown in Table 6. To establish whether the one-factor model functioned significantly worse than the two-factor model, the chi-squared indices were compared between the two models. Since the data is skewed, a Satorra-Bentler scaled chi-square difference test was implemented for this comparison. For the numerical model, the one-factor solution was found to be significantly worse with ${\chi^2}_\Delta(1) = 292.65$ $p \leq 0.0001$. For the categorical scale, the one-factor model also performed significantly worse with ${\chi^2}_\Delta(1) = 88.142$ $p \leq 0.0001$. In addition, the change in the remaining metrics was computed and is shown in Table 6. Based on these results, the one-factor solution is rejected and the two-factor scales are accepted. The *Feature Understandability Scale* is best

Table 4: Feature Understandability Scale for Numerical Items. Factors, Items and factor loadings are displayed alongside item descriptive statistics (mean, standard deviation, corrected item-factor correlation and communalities.

N	Dimension	Item	Factors		M(SD)	r	h^2
			1	2			
1	Understanding & Measurement	I can understand the scale (units) of the feature.	0.63		3.73, (1.03)	0.67	0.47
2	Understanding & Measurement	I can easily understand if a given value of the feature is high or low.	0.66		3.93, (0.92)	0.64	0.46
3	Understanding and Measurement	I know what this feature measures.	0.85		3.96, (0.86)	0.73	0.70
4	Understanding & Measurement	I know what this feature represents.	0.83		4.03, (0.83)	0.73	0.68
5	Understanding & Measurement	I think that I can easily access a definition of the feature.	0.78		3.92, (0.92)	0.64	0.55
6	Feature-Outcome Relation	In my opinion, the feature should be used to predict the outcome.		0.88	3.51, (1.10)	0.67	0.75
7	Feature-Outcome Relation	I think that the feature is important for the outcome.		0.85	3.80, (0.99)	0.69	0.74
8	Feature-Outcome Relation	I think it is fair that the feature influences the outcome.		0.87	3.56, (1.10)	0.69	0.76

Table 5: Feature Understandability Scale for Categorical Items. Factors, Items and factor loadings are displayed alongside item descriptive statistics (mean, standard deviation, corrected item-factor correlation and communalities).

N	Dimension	Item	Factors		M(SD)	r	h^2
			1	2			
1	Understanding & Measurement	I can easily understand how the categories were assessed.	0.87		3.30, (1.22)	0.81	0.74
2	Understanding & Measurement	I can easily understand the order of categories.	0.87		3.49, (1.14)	0.77	0.68
3	Understanding & Measurement	I think it is feasible for me to verify the specific category of the feature.	0.67		3.49, (1.08)	0.77	0.61
4	Understanding & Measurement	I understand all possible values of the categorical feature.	0.80		3.34, (1.20)	0.71	0.58
5	Understanding & Measurement	I know what this feature represents.	0.78		3.56, (1.16)	0.80	0.67
6	Understanding & Measurement	I require no support to understand the feature.	0.73		3.62, (1.12)	0.78	0.63
7	Feature-Outcome Relation	In my opinion, the feature should be used to predict the outcome.		0.79	3.42, (1.01)	0.66	0.63
8	Feature-Outcome Relation	I think that the feature is important for the outcome.		0.74	3.80, (0.92)	0.64	0.58
9	Feature-Outcome Relation	I think it is fair that the feature influences the outcome.		0.87	3.54, (0.98)	0.70	0.74

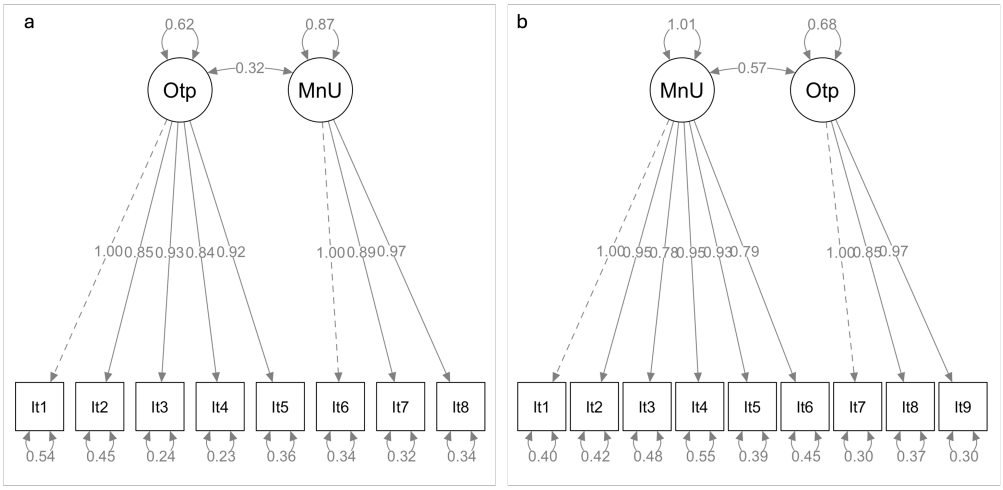


Figure 4: Result of the CFA for the a) Numerical Scale b) Categorical Scale. The Abbreviations are as follows: Otp (Feature-Outcome Relation) and MnU (Understanding and Measurement).

constructed with two factors, with 8 items for numerical features and 9 for categorical features. The two factors represent (1) the understanding and measurement of the feature and (2) the relation of the feature to the predicted output.

For the final scales, several metrics were computed, which are presented in Table 7. McDonald's omega was computed to assess the composite reliability. McDonald's omega is considered acceptable at a value greater than 0.7. The internal consistency was computed using the average variance extracted (AVE), which is considered acceptable at AVE ≥ 0.5 and

Table 6: Results of the 2-factor and 1-factor analysis of the CFA alongside the changes in performance metrics.

		χ^2	χ^2 df	χ^2 p	CFI	TLI	RMSEA	SRMR
Numerical	2-F	122.92	19	$p \leq 0.001$	0.97	0.96	0.081	0.032
	1-F	1095.59	20	$p \leq 0.001$	0.68	0.55	0.260	0.150
Δ					0.09	0.13	0.097	0.049
Categorical	2-F	63.66	26	$p \leq 0.001$	0.99	0.98	0.051	0.028
	1-F	431.92	27	$p \leq 0.001$	0.88	0.83	0.160	0.081
Δ					0.11	0.15	0.110	0.043

Table 7: Metrics for the final scale after CFA.

	Factors	Mc Donald's ω	AVE	F1	F2
Numerical	F1	0.87	0.58	1	
	F2	0.88	0.70	0.44	1
Categorical	F1	0.92	0.65	1	
	F2	0.85	0.65	0.68	1

good at ≥ 0.6. The McDonald's ω showed that both scales have excellent composite reliability. The AVE furthermore indicated that the internal consistency of the factors ranges from acceptable to good for both scales. Finally, the factors within the scales were strongly correlated for the categorical scale and moderately correlated for the numerical scale [48].

5. Conclusion and Future Works

The *Feature Understandability Scale* measures participants' understanding of features in a tabular dataset. The ultimate goal of the scale was to quantify feature understandability for applications in explainable ML and interpretability by design. The initial scale was drafted based on the literature and refined through expert interviews, focus groups and a pilot study. Following the steps laid out by Carpenter [10], an exploratory factor analysis was then conducted. Two two-factor scales, one for numerical items and one for categorical items, were extracted. To confirm this structure, a confirmatory factor analysis was conducted on an independent dataset. The two-factor structure was confirmed, showing significant improvement over a one-factor solution. The factors of both scales were named according to the same logic, with the first factor, 'Understanding and Measurement', containing items that assessed whether users understood the concept underlying the feature, as well as how the value of the feature was measured. The second factor, 'Feature-Outcome Relation', contains the same items for both scales, encompassing how the participant perceives the feature as a predictor for the outcome. The final scale showed good composite reliability and internal consistency. Overall, this manuscript successfully developed a scale to measure the understandability of features, marking an important first step to a new method for interpretability by design.

Given the multi-disciplinary nature of this work, several approaches to future research should be considered. From a psychometric perspective, validation of the scale structure in new samples should be prioritised to elucidate whether the structure is replicable. Additionally, it may be of interest to stratify the sample by gender, ethnicity or level of education to test whether the current scale structure is an artefact of sample fluctuations. While efforts were made to mitigate these effects in the current study through pseudo-random sampling and a large sample size, they could not be entirely negated. In addition to scale replication, translation to other languages and data topics may be of interest. While translation into other languages would permit insight into whether the scale structure is an artefact of the English language, translation to other domains, such as measuring understanding in a language acquisition context, could open doors for more widespread adaptation.

From an ML perspective, the *Feature Understandability Scale* marks the first step towards integrating 'understanding' scores into the ML pipeline, allowing for the favouring of understandable features during the training process and the creation of more understandable explanations. To test whether the scale can improve explanations, future research should aim to integrate it into the ML pipeline, favouring variables with high understandability scores during the explanation process and evaluating the quality of the explanations through a user study. It may furthermore be of interest to assess where in the ML workflow scores are best integrated to improve explanations. While integrating the scores before or in the ML algorithm would permit for an interpretability-by-design solution, in which more understandable features are favoured during the training process, this may present a trade-off with accuracy, as the more understandable features are not necessarily the most informative ones. On the other hand, the scores could serve as a filter method after ML training and before creating an explanation, where relevant features are refined and only understandable features are included in the final explanation. While this would negate the accuracy-explainability trade-off, it would reduce the authenticity of the explanation, as not all relevant features may be included. In addition, it should be assessed whether the use of understandable features consistently leads to the creation of more understandable explanations, or whether other moderators, such as domain, education level, or explanation length, need to be considered.

Despite best efforts, certain limitations of this research need to be ac-

knowledged. First, while pseudo-random sampling was employed through the platform 'Prolific', a true-random sample could not be collected; hence, there remains a chance that the current sample structure is an artefact of sample demographics. Second, despite following the recommendations for best practices set out by Carpenter [10], a possibility remains that items were missed during the initial creation of the scale, and hence the final scale may be incomplete. As such, future research should aim to replicate both the final scale structure and the scale construction process to ascertain whether any relevant considerations were not included in the current research.

The study is the first step toward an explainability-by-design approach that quantifies the user's domain knowledge and enables its integration into the ML workflow. This enables the co-optimisation of accuracy and explainability for tabular data, producing explanations that are understandable by default and enhancing the transparency and auditability of ML models. While future research should aim to replicate and integrate the existing scale, this article marks an important first step in a new approach to explainability-by-design solutions.

Funding and Acknowledgements

This work was conducted with the financial support of Research Ireland - Taighde Éireann, under Grant Nos. 18/CRT/6223 and 12/RC/2289-P2, which are co-funded under the European Regional Development Fund. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

6. Data Availability Statement

All relevant data and code can be found at `github.com/ncrossberg/Feature-Understandability-Scale`. The data is completely anonymised and shared in line with ethics guidelines.

7. Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A.

The definitions provided alongside each draft of the scale are the following:

1. **Feature** - A column in a dataset; e.g. 'Debt' to predict Mortgage
2. **Value** - a value that a feature can take; e.g. -€500 as a value of 'debt'
3. **Categorical feature** - a feature with distinct categories; e.g. phone brands
4. **Measuring Scale** - a method used to assess a numerical feature; e.g. Celsius is a measuring scale for temperature
5. **End Points** - The minimum and maximum values a numerical feature can reasonably take; e.g. the age range of loan applicants may be expected to vary between 18 and 100 years.

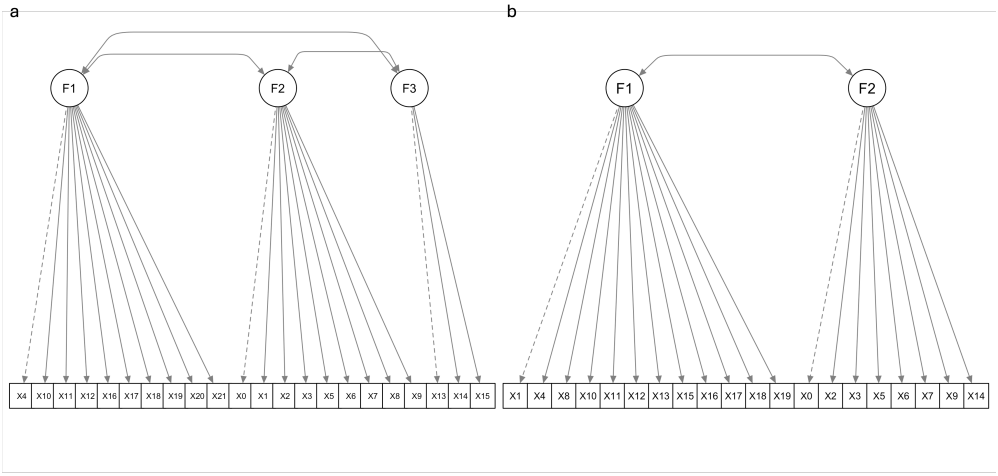


Figure A.5: The assignment of the items to factors after the initial EFA for the a) Numerical Scale and b) Categorical Scale.

Table A.8: Raw and Corrected Pearson's r for both scales before EFA

	Numerical		Categorical	
	Raw	Corrected	Raw	Corrected
X0	0.62	0.6	0.69	0.69
X1	0.71	0.7	0.82	0.81
X2	0.39	0.35	0.49	0.47
X3	0.68	0.67	0.66	0.66
X4	0.34	0.28	0.43	0.38
X5	0.56	0.53	0.57	0.55
X6	0.7	0.69	0.66	0.66
X7	0.68	0.67	0.69	0.69
X8	0.67	0.65	0.76	0.74
X9	0.63	0.6	0.72	0.7
X10	0.72	0.72	0.83	0.82
X11	0.75	0.75	0.78	0.76
X12	0.67	0.66	0.78	0.77
X13	0.7	0.69	0.74	0.72
X14	0.58	0.55	0.75	0.75
X15	0.68	0.67	0.81	0.8
X16	0.69	0.68	0.83	0.82
X17	0.71	0.71	0.79	0.77
X18	0.7	0.7	0.8	0.78
X19	0.68	0.67	0.86	0.85
X20	0.65	0.64	N/A	N/A
X21	0.76	0.76	N/A	N/A

Table A.9: Item-Item Correlation Matrix for the Numerical Scale before EFA.

	X0	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21
X0	1																					
X1	0.68	1																				
X2	0.2	0.23	1																			
X3	0.61	0.68	0.23	1																		
X4	0.11	0.1	0.12	0.11	1																	
X5	0.57	0.54	0.2	0.5	0.23	1																
X6	0.67	0.7	0.26	0.74	0.12	0.54	1															
X7	0.56	0.66	0.25	0.75	0.12	0.47	0.75	1														
X8	0.55	0.6	0.23	0.53	0.24	0.52	0.57	0.54	1													
X9	0.41	0.45	0.22	0.5	0.22	0.38	0.5	0.52	0.46	1												
X10	0.29	0.38	0.25	0.35	0.18	0.25	0.35	0.36	0.38	0.43	1											
X11	0.3	0.42	0.23	0.36	0.2	0.28	0.36	0.36	0.4	0.42	0.82	1										
X12	0.29	0.35	0.25	0.3	0.19	0.24	0.34	0.28	0.36	0.3	0.56	0.57	1									
X13	0.32	0.37	0.24	0.31	0.19	0.26	0.34	0.31	0.38	0.36	0.58	0.61	0.64	1								
X14	0.27	0.3	0.23	0.26	0.16	0.24	0.28	0.2	0.3	0.24	0.39	0.45	0.43	0.57	1							
X15	0.26	0.34	0.23	0.32	0.2	0.24	0.32	0.29	0.35	0.32	0.52	0.56	0.49	0.56	0.71	1						
X16	0.3	0.4	0.25	0.5	0.21	0.25	0.43	0.51	0.35	0.48	0.5	0.5	0.41	0.45	0.36	0.48	1					
X17	0.29	0.35	0.25	0.31	0.23	0.26	0.33	0.32	0.38	0.36	0.56	0.55	0.55	0.51	0.34	0.49	0.55	1				
X18	0.33	0.4	0.26	0.31	0.19	0.28	0.36	0.34	0.36	0.36	0.52	0.53	0.54	0.47	0.33	0.47	0.48	0.71	1			
X19	0.24	0.35	0.18	0.27	0.22	0.21	0.26	0.28	0.35	0.31	0.55	0.6	0.51	0.5	0.4	0.51	0.45	0.6	0.61	1		
X20	0.23	0.3	0.21	0.24	0.21	0.2	0.25	0.28	0.29	0.31	0.5	0.53	0.5	0.47	0.32	0.47	0.44	0.61	0.62	0.66	1	
X21	0.32	0.43	0.22	0.39	0.19	0.26	0.39	0.4	0.37	0.42	0.59	0.64	0.53	0.54	0.36	0.52	0.59	0.65	0.66	0.69	0.71	1

Table A.10: Item-Item Correlation Matrix for the Categorical Scale before EFA.

	X0	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19
X0	1																			
X1	0.56	1																		
X2	0.38	0.35	1																	
X3	0.53	0.56	0.39	1																
X4	0.22	0.28	0.2	0.21	1															
X5	0.51	0.42	0.32	0.38	0.26	1														
X6	0.59	0.54	0.43	0.6	0.18	0.47	1													
X7	0.54	0.57	0.36	0.68	0.2	0.4	0.65	1												
X8	0.54	0.63	0.29	0.46	0.34	0.46	0.48	0.47	1											
X9	0.52	0.52	0.37	0.51	0.23	0.45	0.51	0.53	0.52	1										
X10	0.47	0.67	0.29	0.45	0.36	0.37	0.4	0.48	0.64	0.55	1									
X11	0.47	0.6	0.33	0.38	0.33	0.36	0.41	0.41	0.57	0.5	0.72	1								
X12	0.44	0.61	0.35	0.48	0.32	0.4	0.46	0.49	0.57	0.53	0.66	0.64	1							
X13	0.46	0.55	0.25	0.39	0.3	0.35	0.35	0.38	0.55	0.49	0.65	0.65	0.56	1						
X14	0.53	0.61	0.37	0.6	0.24	0.4	0.53	0.62	0.51	0.55	0.59	0.53	0.59	0.5	1					
X15	0.5	0.67	0.31	0.41	0.32	0.39	0.44	0.46	0.61	0.52	0.71	0.64	0.62	0.63	0.54	1				
X16	0.52	0.72	0.31	0.44	0.3	0.43	0.46	0.48	0.6	0.54	0.69	0.65	0.63	0.63	0.57	0.78	1			
X17	0.43	0.62	0.25	0.38	0.33	0.31	0.35	0.42	0.59	0.53	0.73	0.67	0.63	0.68	0.53	0.69	0.7	1		
X18	0.48	0.66	0.38	0.45	0.3	0.38	0.44	0.47	0.54	0.57	0.68	0.63	0.63	0.56	0.58	0.64	0.69	0.65	1	
X19	0.53	0.69	0.39	0.52	0.28	0.41	0.51	0.53	0.6	0.57	0.75	0.69	0.67	0.66	0.62	0.72	0.74	0.74	0.71	1

Table A.11: Final Item-Item Correlation Matrix for the Numerical Scale. Note that item numbering was reset for the final scale.

		Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
Item 1	I can understand the scale (units) of the feature.	1							
Item 2	I can easily understand if a given value of the feature is high or low.	0.62	1						
Item 3	I know what this feature measures.	0.58	0.55	1					
Item 4	I know what this feature represents.	0.54	0.54	0.73	1				
Item 5	I think that I can easily access a definition of the feature.	0.59	0.55	0.63	0.6	1			
Item 6	In my opinion, the feature should be used to predict the outcome.	0.27	0.23	0.26	0.3	0.29	1		
Item 7	I think that the feature is important for the outcome.	0.3	0.28	0.31	0.36	0.31	0.7	1	
Item 8	I think it is fair that the feature influences the outcome.	0.28	0.22	0.26	0.31	0.3	0.72	0.69	1

Table A.12: Final Item-Item Correlation Matrix for the Categorical Scale. Note that item numbering was reset for the final scale.

	Item	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9
Item 1	I can easily understand how the categories were assessed	1								
Item 2	I can easily understand the order of categories.	0.72	1							
Item 3	I think it is feasible for me to verify the specific category of the feature.	0.63	0.6	1						
Item 4	I understand all possible values of the categorical feature.	0.67	0.68	0.58	1					
Item 5	I know what this feature represents.	0.7	0.7	0.62	0.65	1				
Item 6	I require no support to understand the feature.	0.64	0.62	0.58	0.59	0.66	1			
Item 7	In my opinion, the feature should be used to predict the outcome.	0.46	0.38	0.51	0.43	0.46	0.46	1		
Item 8	I think that the feature is important for the outcome.	0.43	0.4	0.46	0.4	0.43	0.44	0.64	1	
Item 9	I think it is fair that the feature influences the outcome.	0.48	0.45	0.5	0.43	0.47	0.46	0.69	0.61	1

References

[1] Abbey, J. D. and Meloy, M. G. (2017). Attention by design: Using attention checks to detect inattentive respondents and improve data quality. *Journal of Operations Management*, 53:63–70.

[2] Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160.

[3] Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish journal of emergency medicine*, 18(3):91–93.

[4] Anthoine, E., Moret, L., Regnault, A., Sébille, V., and Hardouin, J.-B. (2014). Sample size used to validate a scale: a review of publications on newly-developed patient reported outcomes measures. *Health and quality of life outcomes*, 12(1):2.

[5] Antoniadi, A. M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. A., and Mooney, C. (2021). Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences*, 11(11):5088.

[6] Band, S. S., Yarahmadi, A., Hsu, C.-C., Biyari, M., Sookhak, M., Ameri, R., Dehzangi, I., Chronopoulos, A. T., and Liang, H.-W. (2023). Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. *Informatics in Medicine Unlocked*, 40:101286.

[7] Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M., and Eckersley, P. (2020). Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 648–657.

[8] Brennen, A. (2020). What do people really want when they say they want" explainable ai?" we asked 60 stakeholders. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, pages 1–7.

[9] Brooke, J. (1996). Sus: A 'quick and dirty'usability scale. usability evaluation in industry. pw jordan, b thomas, ba weerdmeester and al mcclelland.

[10] Carpenter, S. (2018). Ten steps in scale development and reporting: A guide for researchers. *Communication methods and measures*, 12(1):25–44.

[11] Chattopadhyay, A., Slocum, S., Haeffele, B. D., Vidal, R., and Geman, D. (2023). Interpretable by design: Learning predictors by composing interpretable queries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7430–7443.

[12] Contreras, J. and Bocklitz, T. (2024). Explainable artificial intelligence for spectroscopy data: a review. *Pflügers Archiv-European Journal of Physiology*, pages 1–13.

[13] Costello, A. B. and Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research, and evaluation*, 10(1).

[14] De Vaus, D. and De Vaus, D. (2013). *Surveys in social research*. Routledge.

[15] Doshi-Velez, F. and Kim, B. (2017). A roadmap for a rigorous science of interpretability. *arXiv preprint arXiv:1702.08608*, 2(1).

[16] Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., and Weisz, J. D. (2021). Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–19.

[17] European Parliament and the Council of the European Union (2024). Regulation (eu) 2024/1689 on harmonised rules on artificial intelligence (artificial intelligence act). *Official Journal of the European Union*. 12.7.2024; Text with EEA relevance; ELI: http://data.europa.eu/eli/reg/2024/1689/oj.

[18] Gorsuch, R. L. (1983). *Factor Analysis*. Psychology Press, 2 edition.

[19] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42.

[20] H., M. Y. (2022). Loan default dataset. Accessed: 2025-06-27.

[21] Halpern, J. Y. and Pearl, J. (2005). Causes and explanations: A structural-model approach. part ii: Explanations. *The British journal for the philosophy of science*.

[22] Hayes, A. F. and Coutts, J. J. (2020). Use omega rather than cronbach's alpha for estimating reliability. but.... *Communication Methods and Measures*, 14(1):1–24.

[23] Hepenstal, S. and McNeish, D. (2020). Explainable artificial intelligence: What do you need to know? In *Augmented Cognition. Theoretical and Technological Approaches: 14th International Conference, AC 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I 22*, pages 266–275. Springer.

[24] Hind, M., Wei, D., Campbell, M., Codella, N. C., Dhurandhar, A., Mojsilović, A., Natesan Ramamurthy, K., and Varshney, K. R. (2019). Ted: Teaching ai to explain its decisions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 123–129.

[25] Hoffman, R. R., Mueller, S. T., Klein, G., Jalaeian, M., and Tate, C. (2023). Explainable ai: roles and stakeholders, desirements and challenges. *Frontiers in Computer Science*, 5:1117848.

[26] Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2018). Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

[27] Holzinger, A., Carrington, A., and Müller, H. (2020). Measuring the quality of explanations: the system causability scale (scs) comparing human and machine explanations. *KI-Künstliche Intelligenz*, 34(2):193–198.

[28] Hong, S. R., Hullman, J., and Bertini, E. (2020). Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–26.

[29] Hu, L.-t. and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, 6(1):1–55.

[30] Islam, M. R., Ahmed, M. U., Barua, S., and Begum, S. (2022). A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Applied Sciences*, 12(3):1353.

[31] Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., and Rosseel, Y. (2025). *semTools: Useful tools for structural equation modeling*. R package version 0.5-7.

[32] Kaiser, H. F. (1974). An index of factorial simplicity. *psychometrika*, 39(1):31–36.

[33] Kharoua, R. E. (2024). Predicting hiring decisions in recruitment data.

[34] Kim, M., Kim, S., Kim, J., Song, T.-J., and Kim, Y. (2024). Do stakeholder needs differ?-designing stakeholder-tailored explainable artificial intelligence (xai) interfaces. *International Journal of Human-Computer Studies*, 181:103160.

[35] Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., and Baum, K. (2021). What do we want from explainable artificial intelligence (xai)?–a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. *Artificial Intelligence*, 296:103473.

[36] Lee, E., Jung, S. Y., Hwang, H. J., Jung, J., et al. (2021). Patient-level cancer prediction models from a nationwide patient cohort: model development and validation. *JMIR Medical Informatics*, 9(8):e29807.

[37] Lewis, D. K. (1986). Causation. *The Journal of Philosophy*.

[38] Liao, Q. V., Gruen, D., and Miller, S. (2020). Questioning the ai: informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–15.

[39] Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., et al. (2024). Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301.

[40] Lopes, P., Silva, E., Braga, C., Oliveira, T., and Rosado, L. (2022). Xai systems evaluation: A review of human and computer-centred methods. *Applied Sciences*, 12(19):9423.

[41] Mahmoud, A. A., Shawabkeh, T. A., Salameh, W. A., and Al Amro, I. (2019). Performance predicting in hiring process and performance appraisals using machine learning. In *2019 10th international conference on information and communication systems (ICICS)*, pages 110–115. IEEE.

[42] Muhammad, D. and Bendechache, M. (2024). Unveiling the black box: A systematic review of explainable artificial intelligence in medical image analysis. *Computational and Structural Biotechnology Journal*.

[43] Namdari, R. (2022). Seer breast cancer data. https://www.kaggle.com/datasets/reihanenamdari/breast-cancer. Kaggle dataset; originally sourced from the 2017 November update of the SEER Program (NCI).

[44] Nwakanma, C. I., Ahakonye, L. A. C., Njoku, J. N., Odirichukwu, J. C., Okolie, S. A., Uzondu, C., Ndubuisi Nweke, C. C., and Kim, D.-S. (2023). Explainable artificial intelligence (xai) for intrusion detection and mitigation in intelligent connected vehicles: A review. *Applied Sciences*, 13(3):1252.

[45] Preece, A., Harborne, D., Braines, D., Tomsett, R., and Chakraborty, S. (2018). Stakeholders in explainable ai. *arXiv preprint arXiv:1810.00184*.

[46] Ribera, M. and Lapedriza, A. (2019). Can we do better explanations? a proposal of user-centered explainable ai. In *Joint Proceedings of the ACM IUI 2019 Workshops*. CEUR Workshop Proceedings.

[47] Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36.

[48] Rottweiler, B. and Gill, P. (2021). Measuring individuals' misogynistic attitudes: development and validation of the misogyny scale. psyarxiv, 1–42.

[49] Saranya, A. and Subhashini, R. (2023). A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. *Decision analytics journal*, 7:100230.

[50] Simkute, A., Luger, E., Jones, B., Evans, M., and Jones, R. (2021). Explainability for experts: A design framework for making algorithms supporting expert decisions more explainable. *Journal of Responsible Technology*, 7:100017.

[51] Sivanathan, D., Bizumic, B., Li, W., and Chen, J. (2024). The unified narcissism scale–revised: Expanding measurement and understanding of narcissism across cultures. *Assessment*, 31(4):839–854.

[52] Tabachnick, B. G., Fidell, L. S., and Ullman, J. B. (2007). *Using multivariate statistics*, volume 5. pearson Boston, MA.

[53] Taherdoost, H., Sahibuddin, S., and Jalaliyoon, N. (2014). Exploratory factor analysis; concepts and theory. *Advances in applied and pure mathematics*, 27:375–382.

[54] Tavakol, M. and Wetzel, A. (2020). Factor analysis: a means for theory and instrument development in support of construct validity. *International journal of medical education*, 11:245.

[55] Teso, S., Alkan, Ö., Stammer, W., and Daly, E. (2023). Leveraging explanations in interactive machine learning: An overview. *Frontiers in Artificial Intelligence*, 6:1066049.

[56] Thompson, B. (2004). Exploratory and confirmatory factor analysis: Understanding concepts and applications. *Washington, DC*, 10694(000):3.

[57] Vilone, G. and Longo, L. (2021). Classification of explainable artificial intelligence methods through their output formats. *Machine Learning and Knowledge Extraction*, 3(3):615–661.

[58] Vilone, G. and Longo, L. (2023). Development of a human-centred psychometric test for the evaluation of explanations produced by xai methods. In *World Conference on Explainable Artificial Intelligence*, pages 205–232. Springer.

[59] Wang, Z., Huang, C., and Yao, X. (2024). A roadmap of explainable artificial intelligence: Explain to whom, when, what and how? *ACM Transactions on Autonomous and Adaptive Systems*, 19(4):1–40.

[60] Weber, P., Carl, K. V., and Hinz, O. (2024). Applications of explainable artificial intelligence in finance—a systematic review of finance, information systems, and computer science literature. *Management Review Quarterly*, 74(2):867–907.

[61] William Revelle (2025). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 2.5.6.

[62] Worthington, R. L. and Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The counseling psychologist*, 34(6):806–838.

[63] Xu, Q., Liu, C., Luo, J., and Liu, F. (2024). Using machine learning to investigate the determinants of loan default in p2p lending: Are there differences between before and during covid-19? *Pacific-Basin Finance Journal*, 88:102550.

[64] Zijlmans, E. A., van der Ark, L. A., Tijmstra, J., and Sijtsma, K. (2018). Methods for estimating item-score reliability. *Applied psychological measurement*, 42(7):553–570.